GNSO IDNs-EPDP Team:
Hybrid Model for String Similarity Review

ICANN

Agenda

- ⦿ Welcome to joint meeting

- ⦿ Why joint session?

- ⦿ ccNSO Perspective on Confusing Similarity and Variants

- ⦿ GNSO Perspective

- ⦿ Wrap-up: further coordination needed?

Why Joint session?

- Board asked ccNSO and GNSO to coordinate their policy development efforts with respect to variant management
  - ccNSO and GNSO appointed liaisons (Anil and Dennis)
  - Comparison of notes
  - Councils meet regularly (twice a year), to oversee coordination

- Variant Management related to confusing similarity validation process (ccPDP4 term) or string similarity review (GNSO EPDP term)

- Both groups made progress in subject area:
  - Opportunity to understand commonalities and differences in approach
  - Opportunity to learn and adjust

ccNSO Perspective on Confusing Similarity and Variants

ccNSO Perspective

- Scope of ccPDP4 is review, overhaul and completion of 2013 recommendations on the selection of IDN ccTLD strings
  - Review: Selection criteria and required documentation

  - Overhaul: Review and adjustment recommendations on confusing similarity validation

  - Completion: include recommendation on definition of variants, variant management, de-selection of IDNccTLD strings and introduce review mechanism(s).

Overhaul Confusing Similarity Validation

- Overhaul 2013 recommendations needed to align with curent procedures under the Fast Track process (which is on-going)

- Updated Procedural perspective

CS Validation Process

CS Evaluation → CS Review → CS Risk Mitigation Appraisal

Overhaul Confusing Similarity Validation: Goal and Criteria

- **Goal.** The goal of the confusing similarity validation is to minimize **the risk to the stability and security of the DNS due to user confusion by exploiting potential visual confusing similarity between domain names**
  - On the risk (based on SAC 060): No-connection may be a nuisance for the user, like a typo, however misconnection may result in the exploitation of the user confusion and this could be avoided though the similarity review.
  - Focus of CS Validation is minimizing risk of misconnection.

- **Standard for evaluation** A selected IDN ccTLD string is considered confusingly similar with one or more other string(s) if the appearance of the selected string in common fonts in small sizes at typical screen resolutions is sufficiently close to one or more other strings so that it is probable that a reasonable Internet user who is unfamiliar with the script would perceive the strings to be the same or confuse one for the other

Overhaul Confusing Similarity Validation: Base of Comparison

A Selected string, and its Requested, Delegatable* Variants should not be confusingly similar with:

o   Any combination of two ISO 646 Basic Version (ISO 646-BV) characters (letter [a-z] codes),

nor

o   Existing TLDs, which includes the already delegated variants or reserved names,

nor

o   Proposed TLDs which are in process of string validation and their requested Delegatable or requested variants (however defined under the ccTLD and gTLD processes)

* Delegatable Variant = Allocatable Variant that is Meaningful representation of name of territory in designated language and script in which designated language is expressed

Overhaul Confusing Similarity Validation: Base of Comparison Rationale

- Goal of minimizing risk of confusing similarity is to minimize risk of **misconnection,** and therefore avoid that a requested CS string is potentially delegated. The goal is **not** to minimize or avoid Non-Connection.

- By definition allocatable variants may be requested at a later stage.

- However, allocatable variants will need to meet all criteria, including confusing similarity and meaningfulness to be delegated.

- By including all allocatable variants in the comparison side, the confusing similarity review could become a reservation system.

GNSO Perspective

Background

EPDP Team focused its discussion on variants' role in the String Similarity Review, but not the other aspects of the review as they are not part of the EPDP charter and already covered by SubPro outputs

Charter Questions

EPDP-IDN Charter asks to **consider any adjustment to the string similarity review due to the variant implementation**: (Charter Question E3)
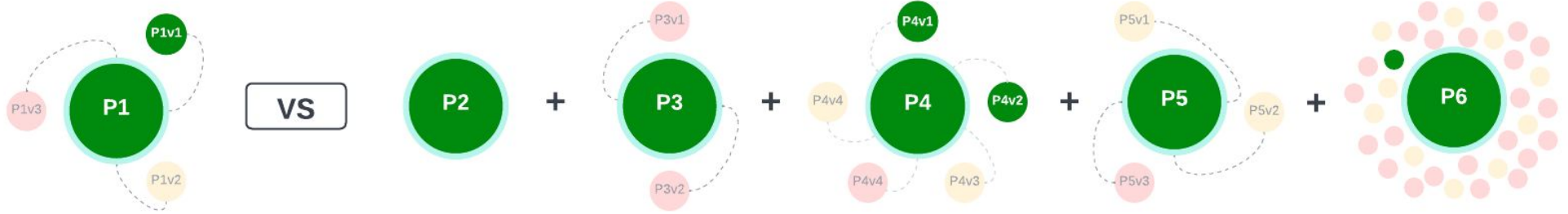
- What role, if any, do the "withheld same entity" variants play? (Charter Question E1)

- What are the potential consequences for the other allocatable variant labels in the same set of a requested variant label, which is rejected as a result of the string similarity review? (Charter Question E3a)
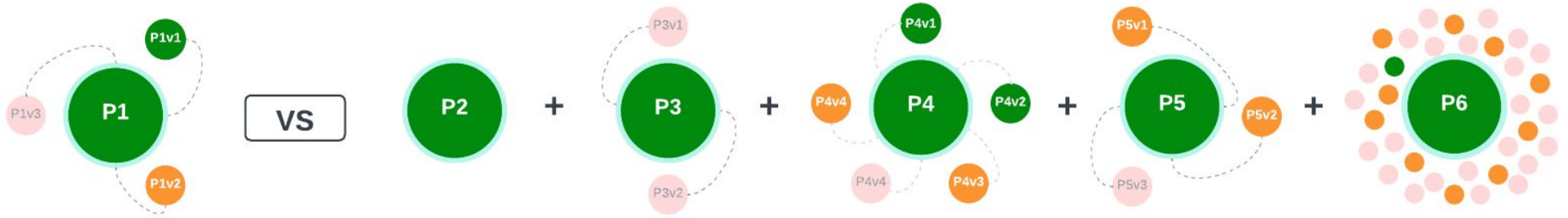
Background (Cont.)

- The EPDP Team began the deliberation by discussing **three (3) possible levels of comparison** among visually confusable strings, as well as analyzed the impact and potential consequences:

    Level 1: Primary + only requested allocatable variants

    Level 2: Primary + all allocatable variants

    Level 3: Primary + all valid variants (blocked + allocatable)

- Following its String Similarity small group's deliberation, the EPDP Team put forward a preliminary recommendation for a *hybrid model, which is a mixed-level approach between level 2 and level 3*

    - *NOTE: Blocked variants of one TLD are NOT be compared against blocked variants of another TLD; everything else is compared against each other*
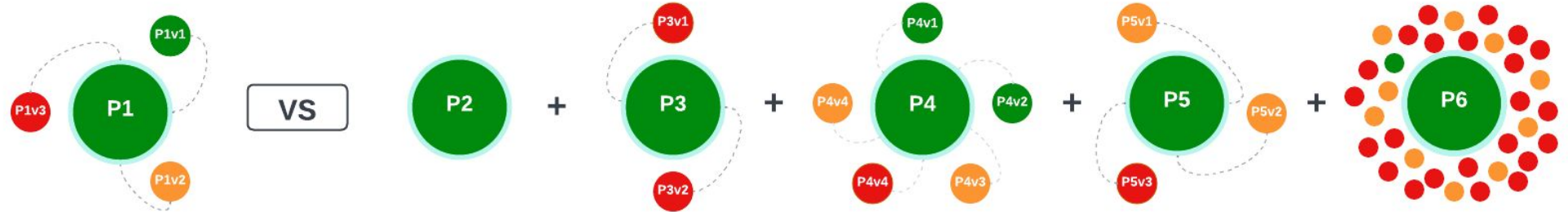
Three Levels of Comparison

Recommendation: Hybrid Model

The string similarity review must be modified to compare:

- **An applied-for primary IDN gTLD and <u>all of its allocatable variant label(s)</u>**

Against:

- **Existing TLDs and <u>all of their allocatable and blocked variant labels</u>;**

- **Strings requested as IDN ccTLDs and <u>all of their allocatable and blocked variant labels</u>;**

- **Other applied-for gTLDs in the same round and <u>all of their allocatable and blocked variant labels</u>;**

- **Reserved Names; and**

- **Any other two-character ASCII strings and <u>all of their allocatable and blocked variant labels</u> (*if the applied-for primary IDN gTLD is a two-character string*)**

Recommendation: Hybrid Model (Cont.)
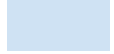
In addition, compare:

- **All of the blocked variant label(s)** of an applied-for primary IDN gTLD

Against:

- **Existing TLDs and all of their allocatable variant labels;**

- **Strings requested as IDN ccTLDs and all of their allocatable variant labels;**

- **Other applied-for gTLDs in the same round and all of their allocatable variant labels;**

- **Any other two-character ASCII strings and all of their allocatable variant labels (*if the applied-for primary IDN gTLD is a two-character string*)**

How Does the Hybrid Model Work: Arabic TLDs Example

Applied-for Primary Strings:

(A1) رکی

(B1) رگے

Allocatable Variants of Primary Strings:

(A2) رکی
(A3) رکی

None
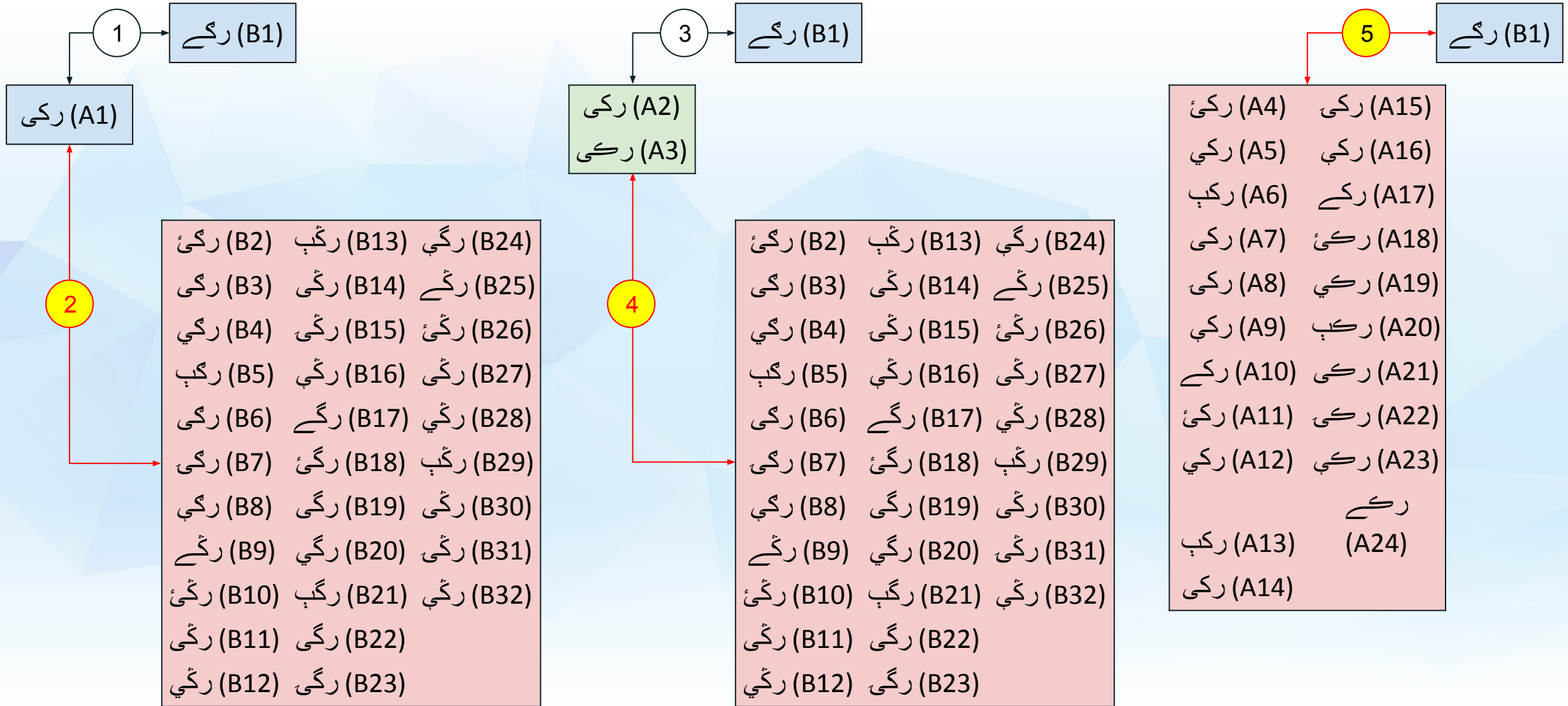
Blocked Variants of Primary Strings:

(A15) رکی	(A4) رکئ	(B13) رگٮ	(B24) رگي
(A16) رکب	(A5) رکي	(B14) رگی	(B25) رگے
(A17) رکے	(A6) رکب	(B15) رگی	(B26) رگئ
(A18) رکئ	(A7) رکی	(B16) رگی	(B27) رگب
(A19) رکي	(A8) رکی	(B17) رگے	(B28) رگي
(A20) رکب	(A9) رکب	(B18) رگئ	(B29) رگب
(A21) رکی	(A10) رکے	(B19) رگی	(B30) رگي
(A22) رکی	(A11) رکئ	(B20) رگي	(B31) رگے
(A23) رکي	(A12) رکي	(B21) رگب	(B32) رگئ
(A24) رکے	(A13) رکب	(B22) رگی	
	(A14) رکی	(B23) رگي	
		(B2) رگئ	
		(B3) رگی	
		(B4) رگي	
		(B5) رگب	
		(B6) رگی	
		(B7) رگی	
		(B8) رگي	
		(B9) رگے	
		(B10) رگئ	
		(B11) رگی	
		(B12) رگي	

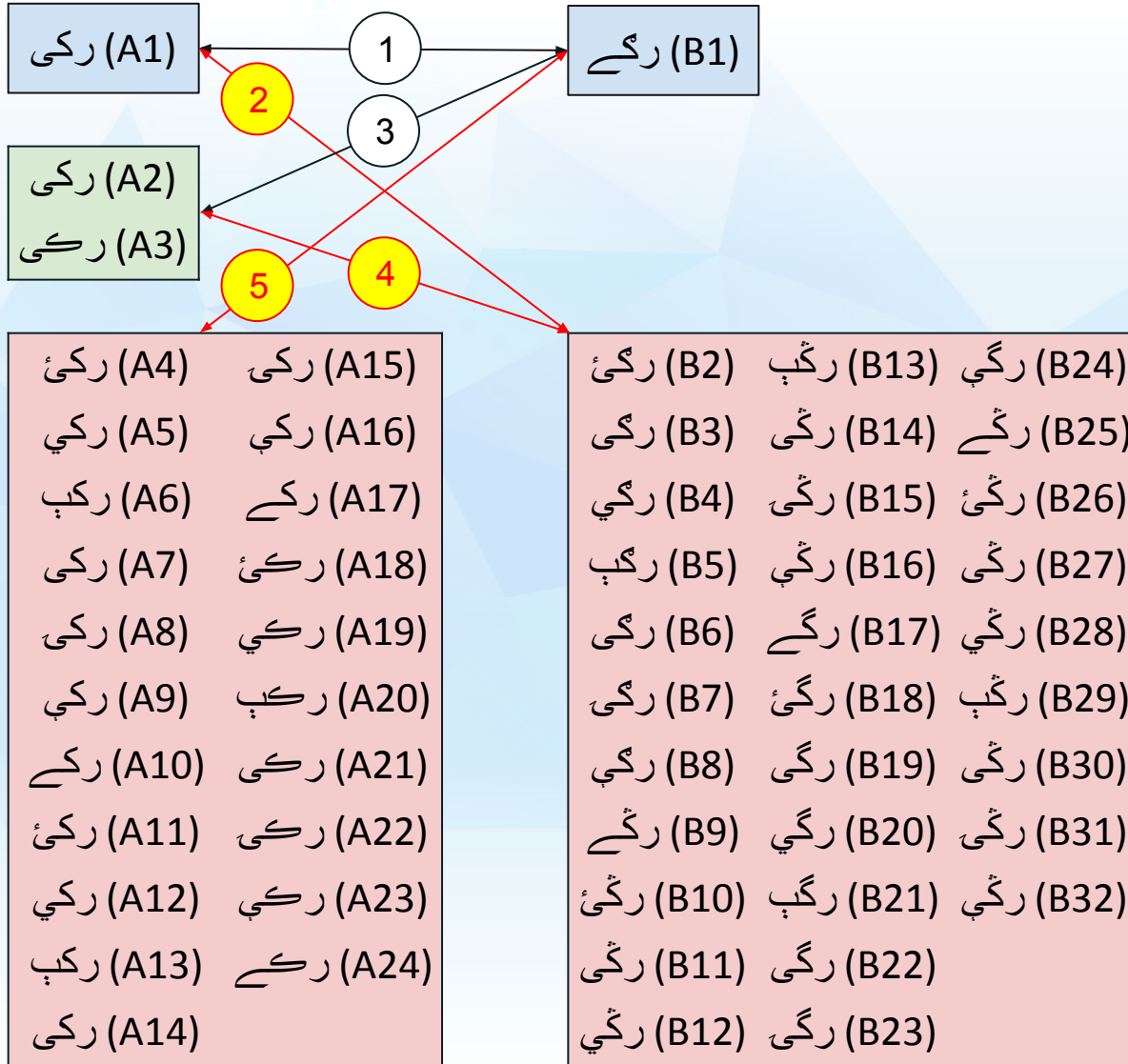# How Does the Hybrid Model Work: Arabic TLDs Example (Cont.)

1 → (B1) رگے

(A1) رکی

2

(B2) رکئ (B13) رگُب (B24) رگي
(B3) رکی (B14) رگی (B25) رگے
(B4) رکي (B15) رگئ (B26) رگی
(B5) رکُب (B16) رگي (B27) رگی
(B6) رکی (B17) رگے (B28) رگي
(B7) رکی (B18) رگئ (B29) رگُب
(B8) رگي (B19) رگی (B30) رگی
(B9) رگے (B20) رگي (B31) رگی
(B10) رگئ (B21) رگُب (B32) رگي
(B11) رگی (B22) رگی
(B12) رگي (B23) رگی

3 → (B1) رگے

(A2) رکی
(A3) رکی

4

(B2) رکئ (B13) رگُب (B24) رگي
(B3) رکی (B14) رگے (B25) رگی
(B4) رکي (B15) رگئ (B26) رگی
(B5) رکُب (B16) رگي (B27) رگی
(B6) رکی (B17) رگے (B28) رگي
(B7) رکی (B18) رگئ (B29) رگُب
(B8) رکي (B19) رگی (B30) رگی
(B9) رگے (B20) رگي (B31) رگی
(B10) رکئ (B21) رگُب (B32) رگي
(B11) رکی (B22) رگی
(B12) رکي (B23) رگی

5 → (B1) رگے

(A15) رکی (A4) رکئ
(A16) رکي (A5) رکي
(A17) رکے (A6) رکُب
(A18) رکئ (A7) رکی
(A19) رکي (A8) رکی
(A20) رکُب (A9) رکي
(A21) رکی (A10) رکے
(A22) رکی (A11) رکئ
(A23) رکي (A12) رکي
(A24) رکے
(A13) رکُب
(A14) رکی

Summary: Arabic TLDs Example

Hybrid model may find the following confusingly similar strings

(2) (B6) رگی & (B3) رگی & (A1) رکی

(4) (B6) رگی & (B3) رگی & (A2) رکی

(4) (B6) رگی & (B3) رگی & (A3) رکے

(5) (A24) رگے & (A17) رکے & (A10) رکے & (B1) رگے

Potential Outcome of the String Similarity Review

رکی (A1) & its variants A2-A24 AND رگے (B1) & its variants B2-B32 get processed in a contention set

If the hybrid model were not used and blocked variants were not taken into account in String Similarity Review

رکی (A1) and رگے (B1) would have been both delegated with the misconnection risk. E.g., a user may mistake رکی (A1) as رگی (B3), a blocked variant of رگے (B1), but arrive at site controlled by a registrant different to رگے (B1).

Rationale for Hybrid Model

The hybrid model should:

- **Meet the singular goal of risk mitigation of failure modes**, which are 1) **denial of service,** and **2) misconnection**

- **Help detect many more pairs of visually confusable strings** and **mitigate confusion risks as much as possible**

- **Reduce computational complexity** by not requiring comparison among blocked variant labels

**Why other levels of comparison may not be sufficiently conservative:**

- Level 1 and 2 may fail to detect some visually confusable strings and increase the risks of failure modes

- Level 3 unnecessarily compares blocked variants against each other with exponential increase of computational complexity

Relevant Background for Consideration of Hybrid Model

❏ **RFC 5891**: Any domain name registry, including that of the root zone, should develop and apply additional **restrictions** as needed **to reduce confusion and other problems** (part of IDNA2008 standard)

❏ **RFC 6921**: Zones higher in the DNS tree tend to have **more restrictive rules**…the context is that the root zone serves the entire Internet population

❏ **SAC089: Confusability cannot be considered in isolation from other issues related to security.** Phishing and other social engineering attacks based on domain name confusion are a security problem for end users

❏ **Staff Paper:** Variant implementation must be done in a way that **operation and maintenance of the DNS not be adversely impacted by the introduction of variants**; it should **avoid including variant TLDs in a manner that would create user vulnerabilities or a probability of confusion**
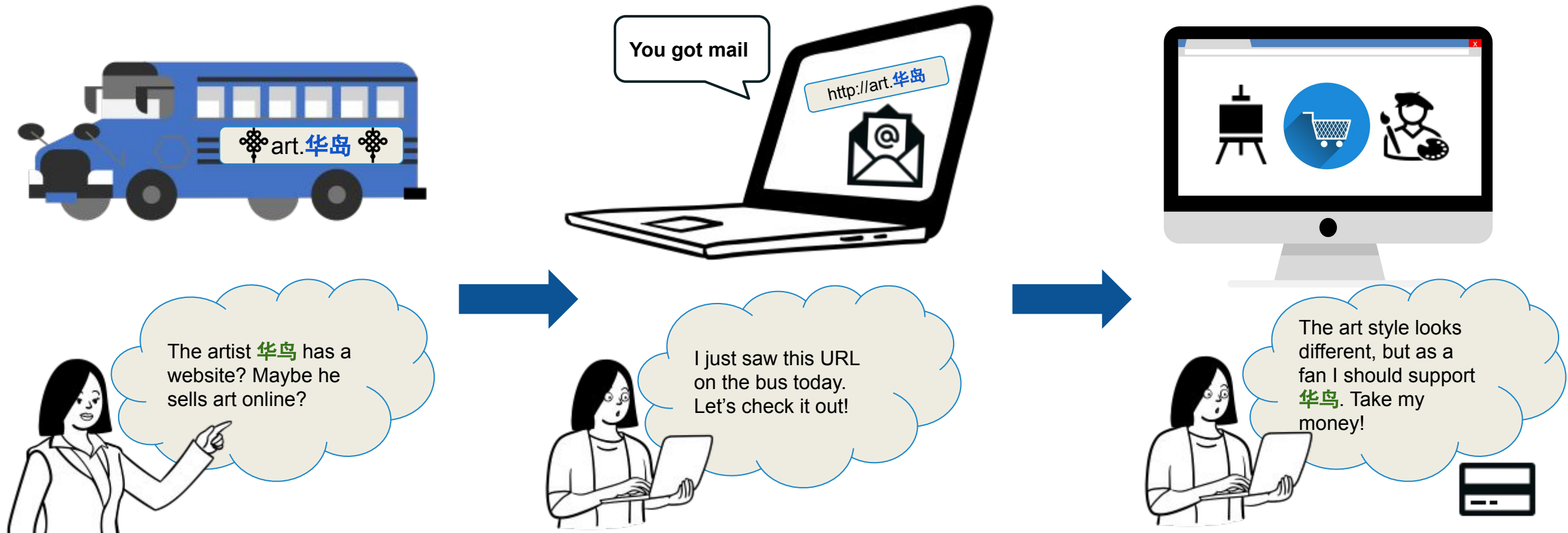
Denial of Service: Example & Illustration

A user attempts to visit http://example.X, reading it as being the same as the http://example.Y that, for example, he or she saw in an advertisement. After typing the address (http://example.X), the connection does not work as http://example.X is not registered.
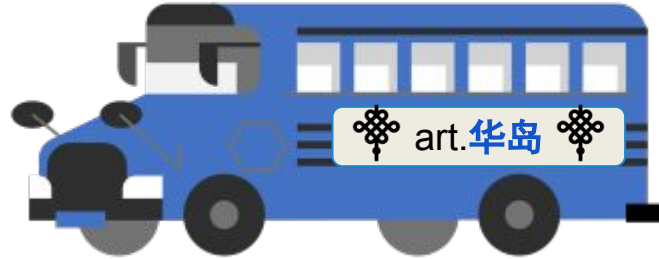
Denial of service will likely cause user confusion and frustration but not harm

Misconnection: Example & Illustration

A user attempts to visit http://example.X, reading it as being the same as the http://example.Y that, for example, he or she saw in an advertisement. After clicking on http://example.Y, the user arrives at a site controlled by a registrant different to http://example.X.

art.华岛

You got mail

http://art.华岛

The artist 华鸟 has a website? Maybe he sells art online?

I just saw this URL on the bus today. Let's check it out!

The art style looks different, but as a fan I should support 华鸟. Take my money!

Misconnection: Potential Consequences

Misconnection may be **more problematic than denial of service** and cause more harm to the user beyond confusion and frustration

Arriving at the wrong site, even legitimate, can **result in credential compromise and accidental exposure of information**

If confusing similarity is maliciously leveraged, it can be a **DNS abuse** vector. When confusion is at the top-level, the possibility of DNS abuse is **much greater** than that at second-level

Why Should Blocked Variants Be Considered?

- A label used in everyday life can be considered a blocked variant label by RZ-LGR calculation

- End users can perceive and intend to access a blocked variant label domain name without knowing that it does not exist in the root zone

Misconnection Involving Blocked Variants: Example and Illustration
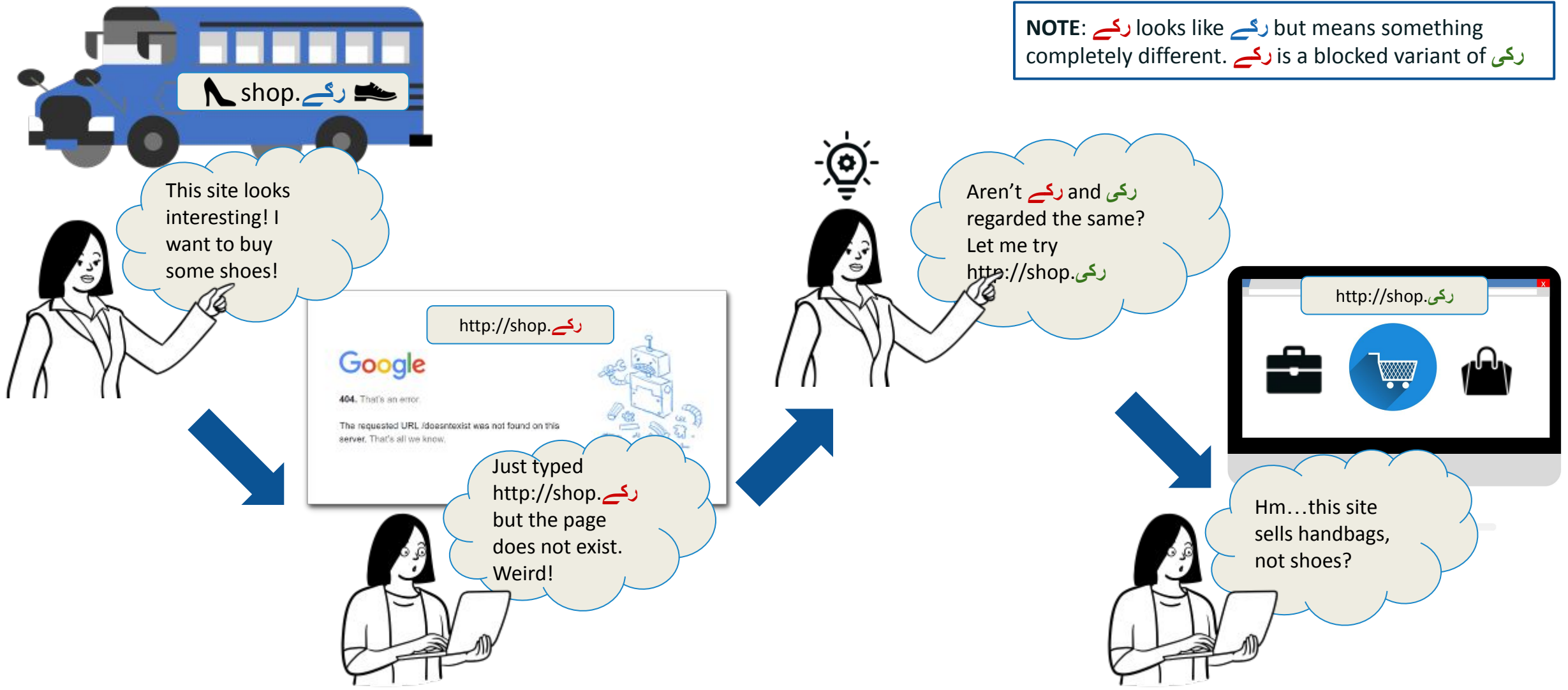
Ongoing Deliberation on Hybrid Model

❏ EPDP Team expressed **general support** for the hybrid model

❏ However, some EPDP Team members expressed **reservations** about including **non-requested allocatable** variant labels and **blocked** variant labels in the String Similarity Review

  ❏ **RySG** raised implementation related suggestions, particularly related to the removal of mix-script labels from further consideration

  ❏ **IPC** suggested an **exception process** for brands to overcome any potential challenges stemmed from the hybrid model

❏ EPDP Team requested **ICANN org to provide operational input to help analyze the implementation complexity** of the hybrid model

❏ Additional Considerations:

  ❏ While the pool of strings that needs to be considered will be large, **language experts in the String Similarity Review panel can evaluate the strings on a case-by-case basis**

  ❏ After the evaluation completes, there are **other mechanisms in the New gTLD Program** – e.g., limited appeal mechanism and objection processes – to review the string similarity panel's decision

Agenda

◉ Welcome to joint meeting

◉ Why joint session?

◉ ccNSO Perspective on Confusing Similarity and Variants

◉ GNSO Perspective

◉ **Wrap-up: further coordination needed?**